A Coding Tool for Multimodal Analysis of Meeting Video

Francis Quek, David McNeill, Travis Rose, and Yang Shi

Vision Interface & Sys. Lab. (VISLab) CSE Dept., Wright State University
Dayton, OH
quek@cs.wright.edu

ABSTRACT

We present our *Visualization for Situated Temporal Analysis*, VisSTA system for coding and analysis of the multimodal communication and interaction in multi-participant meetings. While VisSTA is a much larger system comprising more components for visualizing a variety of data types, we focus on the *Music Score Representation* as the key component for interactive multimodal coding. Although it is not our purpose in this paper to emphasize the theoretical aspects of the social dynamics of free-form brain-storming meetings or of the psycholinguistics of group discourse segmentation, we will motivate our discussion with an example analysis in both dimensions to shot the capacity of VisSTA for such analysis. We described our multi-layered, multi-pass strategy to code the data at the base level, and demonstrated how this coding supports higher level analysis.

VisSTA is capable of handling multiple video/audio sources necessary in the complex task of meeting coding and analysis, and is able to handle long video sequences in excess of the 23-minute microcorpus analyzed. The only limitations are storage and memory for handling the multimedia elements.

INTRODUCTION

We present a system for coding and analysis of the multimodal communication and interaction in multi-participant meetings. This is a complex task in which the speech and behavior of each subject has to be analyzed within modality and across modalities within some theoretical framework of multimodal human communication. Our extended research team comprises researchers in psycholinguistics, psychology, speech and signal processing, computer vision and human-computer interaction researchers. We will motivate our discussion of this coding with an example analysis of 30-seconds of real meeting data. We utilize the *Visualization for Situated Temporal Analysis* (VisSTA) tool (Quek and McNeill 2000; Quek, Shi et al. 2002). We present the theory and implementation of VisSTA elsewhere in this workshop. This paper concentrates on the application of VisSTA for meeting coding and analysis.

THE MEETING DATA

Our data comes from a multi-camera pilot microcorpus collected at the National Institutes for Standards and Technology (NIST) meeting room facility on July 27, 2003. Four individuals (two male, two female) were tasked with choosing the top five news stories for the week of July 22, 2003. They were given 20 minutes to identify the stories from a minimum of 3 categories (out of a set of 5 categorical options), and produce a PowerPoint® slide summarizing their conclusions. Participants were initially seated on opposite sides of a 4ft x 9ft table. They were provided with access to a whiteboard, and a computer equipped with a

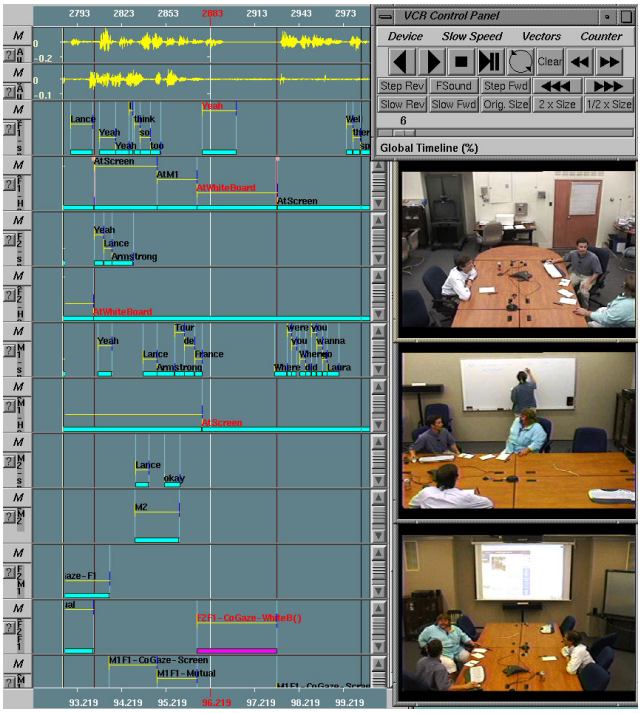

Figure 1. The VisSTA Coding Interface

web browser. The computer display was projected at the front of the room. Five cameras were used, four of which were placed approximately five feet from the middle of each edge of the table. The fifth camera provided a close-up view of the whiteboard. A sixth video recorded the contents of the computer screen. Audio was recorded using both fixed distance wireless microphones worn by the speakers and a suite of table-top noise-canceling microphones. Video and audio were time-synchronized using a movie-style clapper.

The subjects were familiar with each other, and they were told to determine their own roles in the meeting. The meeting participants worked quickly and effectively with rich interaction, with various activities including planning, brainstorming, negotiation, and decision-making. The individuals assumed their roles quickly. The subjects can be seen in the video frames on the right of Figure 1. We designate the subjects as follows: M2: (Male 1) seated and operating the keyboard; M2: (Male 2) seated opposite M1; F1: (Female 1) seated next to M1; and F2: (Female 2) originally opposite F1, but she took the role of the 'scribe' and generally stood by the whiteboard. These roles were maintained for the duration of the meeting.

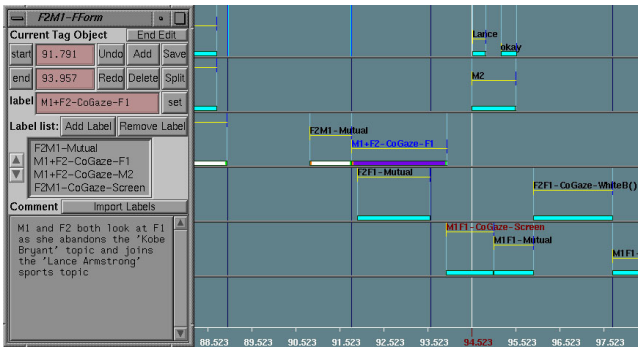

Figure 2. Editing the Music Score

CODING SETUP

We performed a fully manual analysis of 30 seconds of video using VisSTA. Figure 1 shows a condensed view of the VisSTA coding interface configured for this purpose. VisSTA features a flexible interface that supports a variety of visualization and coding, and the user is free to constitute analysis/coding components to suit the purposes of her research.

The right side of Figure 1 shows the *VCR-Style Control Panel*, *VCP*, and a subset of 3 videos from the microcorpus. VisSTA defines a *Current Time Focus*, *CTF* that is common to all components of the system. All three videos and the *Global Timeline* in the *VCP* are kept synchronized with the *CTF*. As the video plays, this synchrony is maintained for all components.

The window left side of Figure 1 is an *Animated Graph Representation Panel*, *AGRP*. An *AGRP* is a container object that defines a time resolution and a *CTF* (represented by the vertical light-colored line running through the middle of the *AGRP*) for all *Animated Graph Panes*, *AGP*s placed in it. As the videos in VisSTA play, the graphs animate like strip charts to maintain *CTF* synchrony. Figure 1 shows two kinds of *AGP*s. The top two *AGP*s are audio signal *AGP*s, and the rest are *Music Score Representation*, *MSR* panes. The latter permits coding and visualization of time-occupying symbolic entities. Each *Score Object* in a *MSR* has a label and may contain other annotations. This is represented visually as a cascading set of lines under the corresponding labels. These labels are user-defined to suit the purposes of the coding. Figure 2 shows an *MSR* editor window in which a *Score Object* (labeled "M1-F2-CoGaze-F1") in the "F2M1-FForm" *MSR* pane is being edited.

*MSR*s can represent any symbolic entity. The choice of semantics of each *MSR* is driven by the analysis at hand. The top 8 *MSR*s in Figure 1 code the words spoken, and the gaze orientations respectively of *F1*, *F2*, *M1*, and *M2* in order. The pair-wise gaze patternings of the subjects are coded in the graphs beneath these within-subject codings (some of these pairwise *MSR* codings are scrolled off the screen).

EXAMPLES OF ANALYSES

We normally process the video to obtain motion traces of the subjects' hands and the subjects' head positions and orientations, and perform a semi-automated forced alignment of each subject's speech (see (Quek, McNeill et al. 2002) for example). Such automated/semi-automated analyses typically help in the coding process. For this microcorpus, the data was of poorer quality (much higher video compression) than we normally employ, and would have necessitated more adjustments than was worth the effort of doing a short coding example. Hence, all the coding described here is done manually using only VisSTA.

BASE CODING

We took a multi-pass multi-layered approach to code the multimodal, multi-participant meeting data. In the first pass, we employed four *MSR* panes to code the speaker turns (there are many instances of overlapping speech in the microcorpus). Next, we instantiated the four 'subject word' *MSR* panes and coded the beginnings and ends of every word. We were assisted in this by the visualization of the audio signal *MSR*s. Also, VisSTA employs the concept of the 'primary video stream' to determine the audio source being played. This stream is chosen by the user and may be changed. This allows us to associate two independent audio streams with each video stream (as stereo components), and to focus on the speech signal from different head-mounted microphones separately.

In the third pass, we coded the gaze orientations of each subject. We observe that subjects tend to gaze at a variety of 'gaze-attractors'. These are the other subjects, the whiteboard, the projection screen, and other artifacts (like clipboards). We estimated the gaze-attractor for each subject manually by studying the multiple video streams. We deemed a gaze at a particular attractor to begin when the subject begins to turn toward that attractor (i.e. at the earliest point where the gaze intention may be detected in the video), and to end when the subject begins to turn toward another attractor. VisSTA's multiple speed playback and single frame stepping capability were critical in making these assessments.

In the fourth pass, we considered the gaze patterning between subjects in a pairwise fashion (e.g. co-gaze at each other, sharing the same gaze-attractor, alternating between gazing at the other subject and a common attractor) in different *MSR*s. Figure 2 shows the coding of a *Score Object* at this pairwise level. This illustrates how VisSTA supports coding and annotation. The user can set the extent of each *Score Object* (the current *Score Object*) in three ways. First, she can click and drag the extrema of its representation in the *MSR* pane. Second, she can move the system to the appropriate *CTF* and click on the *Start* or *End* button on the editing window on the left of Figure 2. We find this to be most effective for detailed coding since we can use the *CTF* the VisSTA animation of the entire dataset for fine time alignment. Third, the user can type the time value directly into the editor window. Beside the *Score Object* label, the user can also enter a free-text comment.

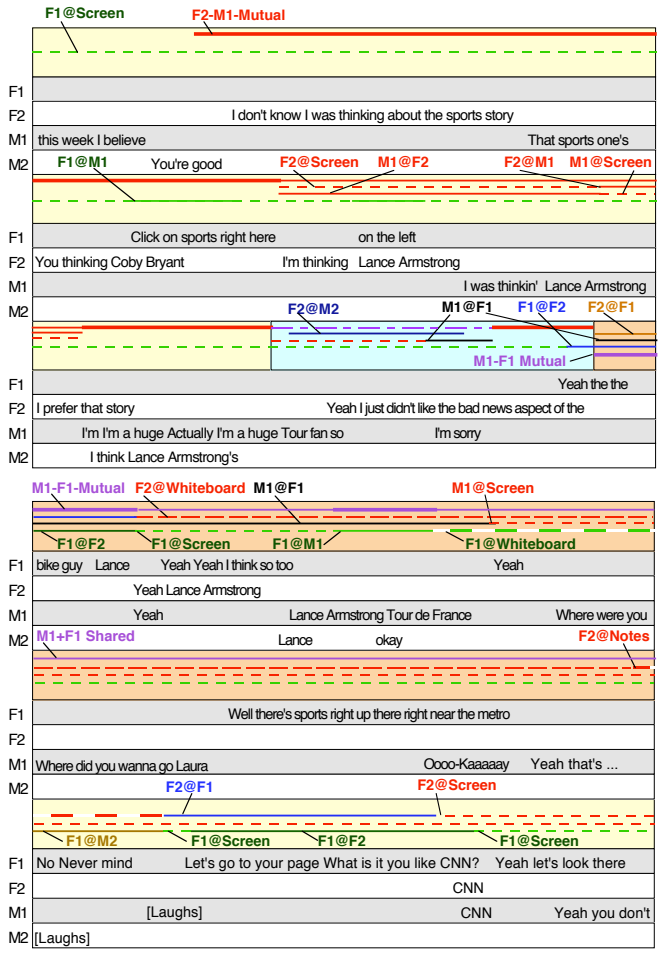

Figure 3. F-Formation coding from NIST meeting video

In the fifth pass, we coded the hand gesture phases of the meeting participants in turn.

ANALYSES

The base coding described above supports two kinds of analysis of our 30-second data segment. In the first analysis, we notice the dominance relationship between *F1* and *F2* from the gaze and speech patterning. In the second analysis, we use the concept of the *F-Formation* to do a high level segmentation of the discourse.

DOMINANCE RELATIONSHIPS

In the discourse segment at hand, the participants discussed the dominant sports story of the week. *F2* advocated the Tour de France victory of Lance Armstrong and *F1* was partial to the breaking scandal story relating to Kobe Bryant. We reproduce the multiple thread speech and gaze coding in a format similar to the VisSTA music score in Figure 3. Each 'staff' of this score contains the speech streams of *F1, F2*, *M1* and *M2* beneath a set of labeled horizontal bars indicating the gaze patterning of the subjects. There are six staves in all.

In the segment spanning the first 3 staves, while *F2* was advocating the Lance Armstrong story, *F1* was trying to get *M1* to select the Kobe Bryant story on the computer. She never uttered the name of the story. In gaze patterning, *F1* and *M1* shared a mutual gaze from up to the middle of the second staff before they engaged in a pattern of alternating gazes: [*M1* at *F2* while *F2* gazes at screen] and [*F2* at *M1* while *M1* gazes at screen]. This sequence continues into the third staff when they reengage in co-gaze. In the middle of staff 3, *F2* glances over at *M2* who was silent throughout, probably to engage him in her story choice. At this point *M1* acknowledges *F1* for the first time (saying "I'm sorry" and looking at her). Immediately, *F2* returns her gaze to *M1* and engages him in a stretch of co-gaze.

This pattern of exclusion of *F1* as a gaze-attractor continued until the end of the third staff when she capitulates and joins the Lance Armstrong story. At this point, *F2* begins looking at *F1* and *F1* and *M1* engages in a period of co-gaze. In the speech coding, one can see a large number of the word "Yeah", as all agreed on the Armstrong story. The new pattern of gaze including *F1* continues through this period as *F2* leaves the field and writes on the whiteboard. A behavioral element not coded in the gaze direction is the posture of *F2*. Each time she engages either *M1* or *M2*, she moves toward him while speaking to him. This triggers an episode of co-gaze.

This pattern of conflict and resolution occurred again is almost exactly the same way (with *F2* engaging *M1* and *M2* and excluding *F1* until *F1* gave up her choice of story) nearly a minute later in the data when the participants were discussing weather-related stories.

F-FORMATIONANALYSIS

In our second analysis, we employed the psycholinguistic device known as the F-Formation first identified by Adam Kendon (Kendon 1977; Kendon 1990). This formulation permits the inference of discourse and interaction structure by observing the maintenance of an 'O-Space' to which two or more individuals have exclusive and shared access. Using this device, we added an additional layer of coding to reflect F-formations.

Figure 3 reproduces the contents of the F-Formation *MSR*s that were coded in VisSTA. The top tier of each staff shows gaze patterns in terms of F-formations. We represent mutual gaze as a thick solid line, gaze at the screen/whiteboard as a dashed line, and gaze at an individual (unshared) as a thin solid line. We highlight different discourse segments uncovered by the F-Formation analysis using different shades. It should be noted that the boundaries of the proposed discourse segments also correspond to shifts in eye gaze.

The first F-Formation involving *F2* and *M1* is light-shaded covering the first two-and-a-half staves. The O-space in this case is the mutual gaze of the principals and the alternating gaze to each other and the screen. Simultaneous to this *F2* speaks to *M1* and holds his attention. During this conversation *F1* attempts to gain control of the floor by directing *M1* to "click on sports"; in the video *F1* is also seen pointing at the screen with her left hand, her arm remaining extended as she looks back to *M1* and at the screen again. She is not included in the F-Formation as no one attends to her.

This F-Formation breaks down in the middle of the third staff. At the end of the third staff, a new F-Formation

involving *F1* and *M1* begins, and stretches to the end of the fifth staff. The essential O-space in this instance is the co-gaze of *F1* and *M1*. As we observed earlier, *F1* had by then abandoned her coup d'etat and is permitted back to the central thread of the discussion, and *F2* having won the argument moves to the whiteboard to record the outcome. This second F-Formation is accompanied by the round of "Yeahs" uttered by all participants. This F-Formation ends at the beginning of staff 6 where the participants proceed to look for a new topic of discussion.

Another approach to the social dynamics of the interaction among participants is to track social gaze – who is being looked at by whom, even briefly. This analysis is distinct from the F-formation analysis in that gaze is analyzed by single individuals, rather than by pairs, and includes only gaze at another person, and excludes gaze at objects in the room. The following table summarizes the directed gaze patterns for 5 minutes of interaction.

Gaze Target

Gaze Source	F2	F1	M1	M2	∑
F2			1	2	3
F1	3		4	3	10
M1	6	1		2	9
M2	3		2		5
∑	12	1	7	7	

Interactive gaze patterns reveal a pattern of social dynamics in the 4-person group during the 5-minute sample. We note that:

- F2 is the most frequent target (at whiteboard).
- F1 is the most frequent gaze source.
- M1 is frequent target at keyboard.
- M2 is as frequent a target as M2, without functional role.

The gaze patterns thus reflect the active roles of M1 (at the keyboard) and F2 (the scribe), but also a seeming pattern of deference to M2, suggesting a figure independently important within the group, though relatively passive in this situation. And finally, the asymmetry of interactions with F1 shown in the F-Formation analysis are reinforced here, in that she was not only the least frequent target of gaze but was also the most frequent source of it.

CONCLUSION

We have presented our VisSTA system for coding and analysis of the multimodal communication and interaction in multi-participant meetings. While VisSTA is a much larger system comprising more components for visualizing a variety of data types, we focused on the *Music Score Representation* as the key component for interactive multimodal coding.

Although it is not our purpose in this paper to emphasize the theoretical aspects of the social dynamics of free-form brain-storming meetings or of the psycholinguistics of group discourse segmentation, we did a example analysis in both dimensions to motivate the capacity of VisSTA to do such analysis. We described our multi-layered, multi-pass strategy to code the data at the base level, and demonstrated how this coding supports higher level analysis.

VisSTA is capable of handling multiple video/audio sources necessary in the complex task of meeting coding and analysis, and is able to handle long video sequences in excess of the 23-minute microcorpus analyzed. The only limitations are storage and memory for handling the multimedia elements.

ACKNOWLEDGMENTS

We thank John Garofolo at NIST for providing access to the meeting room data.

This research has been supported by the U.S. National Science Foundation STIMULATE program, Grant No. IRI-9618887: *Gesture, Speech, and Gaze in Discourse Segmentation*; the NSF KDI program, Grant No. BCS-9980054: *Cross-Modal Analysis of Signal Sense: Multimedia Corpora and Tools for Gesture, Speech, and Gaze Research*; and the NSF ITR program, Grant No. ITR-0219875: *Beyond the Talking Head and Animated Icon: Behaviorally Situated Avatars for Tutoring*; and the Advanced Research and Development Activity ARDA VACEII grant 665661: *From Video to Information: Cross-Model Analysis of Planning Meetings*

REFERENCES

Kendon, A. (1977). "Spatial Organization of Social Encounters: The F-Formation system". *Studies in the Behavior of Social Interaction*. A. Kendon. Lisse, Peter de Ridder Press.

Kendon, A. (1990). **Conducting Interaction: Patterns of Behavior in Focused Encounters**. Cambridge, Cambridge University Press.

Quek, F. and D. McNeill (2000). "A multimedia system for temporally situated perceptual psycholinguistic analysis." *3rd International Conference on Methods and Techniques in Behavioral Research, Measuring Behavior*, Nijmegen, The Netherlands.

Quek, F., D. McNeill, et al. (2002). "Multimodal Human Discourse: Gesture and Speech." *ACM Transactions on Computer-Human Interaction* **9**(3): 171-193.

Quek, F., Y. Shi, et al. (2002). "VisSTA: A Tool for Analyzing Multimodal Discourse Data". *Seventh International Conference on Spoken Language Processing*, Denver, CO.